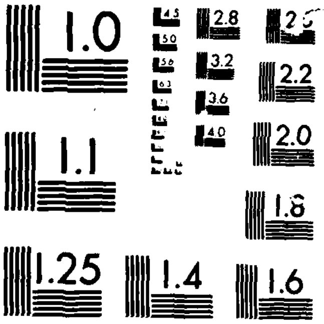
AD-A167 413   OUTER PRODUCT CALCULATIONS(U) AERODYNE RESEARCH INC   1/1
BILLERICA MA J GRUNINGER MAR 86 ARI-RR-526
N00014-84-C-2387

UNCLASSIFIED   F/G 9/2   NL



MICROCOP'   CHART

ARI-RR-526

OUTER PRODUCT CALCULATIONS
ANNUAL REPORT

Prepared by

John Gruninger
Aerodyne Research, Inc.
45 Manning Road
Billerica, MA 01821

Prepared for

Naval Research Laboratory
4555 Overlook Ave., SW
Washington, DC 20375

Prepared Under Contract No. N0014-84-C-2387

March 1986

TABLE OF CONTENTS

Section | Page

INTRODUCTION iv

1 ANALYSIS OF THE EFFECTS OF NOISE ON GENERALIZED INVERSE AND CORRELATION MATRIX ASSOCIATIVE MEMORIES, ... 1-1

2 ADAPTIVE ASSOCIATIVE MEMORY MODELS WITH POTENTIAL OPTICAL IMPLEMENTATION, ... 2-1

3 ON OPTIMAL ASSOCIATIVE RECALL FROM AN INCOMPLETE INPUT VECTOR, and 3-1

4 PROJECTION OPERATORS AND SUBSPACE METHODS OF PATTERN RECOGNITION. 4-1

INTRODUCTION

This report is presented in four sections. In Section 1 an analysis of the effects of noise on associative memories is presented. Special emphasis is given to generalized inverse memories.

In Section 2 of this report, several algorithms which are appropriate for implementation on the NRL spatial light modulator were compiled. These include nonlinear associative memory models and a pseudo inverse memory model that is optimum for incomplete input patterns.

In Section 3 of this report an analysis of a memory model that is optimum in the least squares sense for input patterns with missing components was analyzed. This memory was shown to be derivable from a different optimization principle. This optimization produces that memory which has a minimum average noise output for a given error in recall.

In Section 4 of this report we discuss projection techniques and subspace methods. These approaches will allow the design of dynamic memory and recall schemes that are nonlinear, have local connectivity and can be robust against distorted input patterns. Partial contents:

1. ANALYSIS OF THE EFFECTS OF NOISE ON GENERALIZED INVERSE AND CORRELATION MATRIX ASSOCIATIVE MEMORIES

Associative memory matrices arė constructed from pairs of vectors. Each pair consists of an input vector and an output vector. If the input and output vectors are the same the memory is auto associative. If they are different the memory to hetero associative. The input vectors have ℓ components the output vectors have p components. The number of pairs of vectors, n, is typically less than the number of components. A correlation matrix memory is constructed as

$$M = Y X^T = \sum_{i}^{n} Y_i X_i^T$$

the outer product of the n pairs. A generalized inverse associative memory is constructed as

$$M = Y X^I = \sum_{i}^{n} Y_i X_i{}^I$$

where X^I is a generalized inverse of X. The memories have dimension $p \times \ell$.

There are several properties of true inverses which carry over to generalized inverses. Four have been used to define useful classes of generalized inverses.[1]

The four properties are:

$$A A^I A = A \qquad (1)$$

$$A^I A A^I = A^I \tag{2}$$

$$(A A^I)^T = A A^I \tag{3}$$

$$(A^I A)^T = A^I A \tag{4}$$

We use the notation $A^{(i,j,k,l)}$ to indicate which of the four properties are satisfied (for example, A^1 satisfies (1), $A^{1,4}$ satisfies 1 and 4). The Moore Penrose pseudo inverse satisfies all four.

$$A^+ = A^{1,2,3,4}$$

Associative memories using the Moore Penrose pseudo inverse have been proposed and studied by Kohonen.[2] Any inverse which satisfies (1) provides a solution to $AZ = b$. $Z_0 = A^1 b$ is a particular solution and $Z = A^1 b + (1-A^1 A)y$ for arbitrary y is a general solution.

Any inverse which satisfies (1) can be used to generate a memory matrix with no crosstalk.

Let

$$M = YX^1$$

Then M operating on an exact copy of an input X_i produces a correct copy of the output Y_i. In correlation matrix memories the cross talk is a function of the overlap of inputs. Since generalized inverse associative memories generate no crosstalk the output $\hat{Y}_k$ for an input $\hat{X}_k$ with additive noise is

$$M \hat{X}_k = \hat{Y}_k$$

where

$$\hat{X}_k = X_k + N_k$$

and

$$\hat{Y}_k = Y_k + N_k^o$$

N_k and $N_k{}^o$ are the input and output noise respectively. We analyze here the role that the inputs and outputs have on the signal to noise ratios. Without loss of generality we can assume that both input and output vectors are normalized to unity. If inputs and outputs are all normalized to the same length, then the ratio of the output noise to input noise is equal to the ratios of input to output signal to noise for the generalized inverse memories. For correlation matrix memories, crosstalk contributes to total input noise. Since $M\ (X + N) = Y + N^o + C$, the more telling ratio is

$$\frac{\| N^o + C \|}{\| N \|}$$

for correlation memories. The strength of the output noise versus the input noise is given by

$$\frac{\| N^o \|^2}{\| N \|^2} = \frac{\| M\ N \|^2}{\| N \|^2} = \frac{N^T M^T M\ N}{N^T N}$$

This ratio is bounded by the maximum and minimum eigenvalues of $M^T M$ that is

$$\tau(\mu_{min})\ \| N \|^2 \leq \| N^o \|^2 \leq (\mu_{max})\ \| N \|^2$$

The output noise will be largest along the direction of the eigenvector of M^TM associated with μ_{max} and smallest along the direction of the eigenvector associated with μ_{min}. If the input noise is white the average output noise to input noise can be obtained by averaging the ℓ eigenvalues of M^TM. This average is equal to the trace divided by the number of components.

$$\|N^o\|^2 = \frac{1}{\ell} T_R \, (M^TM) \; \|N\|^2$$

If there are many zero eigenvalues, this average will be small. Indeed typically the number of components in the input vectors, ℓ, is large compared with the number of pairs, n, and the rank of M^TM will be less than or equal to n, depending on the linear independence of the inputs. If the inputs are linearly independent there will be n nonzero eigenvalues and the trace can be replaced by n times the average of these n largest eigenvalues, $\hat{\mu}$

$$\|N^o\|^2 = \left(\frac{n}{\ell}\right) \hat{\mu} \; \|N^2\| \qquad .$$

The n/ℓ reduction in noise has been derived before for pseudoinverse auto-associative memories.[2-3] The n/ℓ reduction is obtained by any correlation matrix memory or generalized inverse memory. The memory need not be auto-associative nor even square. The reduction depends on the number of components in the input vectors and is independent of the number of output components. Increasing the number of components in the input vectors for the sake of noise reduction alone is not advised however, as the noise that is reduced is noise that is perpendicular to the space of the inputs and which does not contribute to the confusion among inputs. The input noise is a combination of perpendicular and parallel noise.

$$N = N_{\parallel} + N_{\perp}$$

These contributions are the contributions in the subspace associated with the inputs, $N_{\parallel}$, and contributions which are in the complement subspace. The noise in the complement subspace is orthogonal to the subspace associated with the inputs. For correlation matrix memories $X_i^T N_{\perp} = 0$ for all stored inputs. For generalized inverse associative memories $X_i^I N_{\perp} = 0$ for all stored inputs and therefore

$$MN = MN_{\parallel}$$

$$MN_{\perp} = 0$$

$N_{\parallel}$ can be expressed as a linear combination of the stored inputs since the inputs form a basis

$$N_{\parallel} = \sum_{i}^{n} X_i \gamma_i$$

Increasing the number of input components (ℓ) on the surface appears to be helpful in reducing output noise. However unless the added components change the eigenstructure of $M^T M$ and reduce the crosstalk, there is no reduction in confusion. When the input noise is white it is partitioned equally in all directions on the average with average total strength of σ^2. The partitioned between the parallel and perpendicular subspace depends only on their relative dimensions

$$\| N_{\parallel} \|^2 = \frac{n}{\ell} \sigma^2$$

and

$$\| N_{\perp} \|^2 = \frac{\ell - n}{\ell} \sigma^2$$

It is only parallel input noise that mixes inputs and causes confusion. The average output noise strength depends on the average nonzero eigenvalue of M^TM and the average parallel input noise strength

$$\|N^o\|^2 = \hat{\mu} \quad \|N_\|\|^2$$

The output noise strength in general is given by

$$\|N^o\|^2 = \|MN_\|\|^2$$

and the covariance matrix for the output noise can be found from the covariance matrix of the input noise Let R_N be the covariance matrix of the input noise. Then $R_N o = MR_NM^T$ is the covariance matrix of the output noise. If the input noise is white, the output covariance is proportional to the outer product of the memory with itself $R_N o = \sigma^2MM^T$. The covariances and strengths of the output noise is a function of the singular values of the memory matrix or equivalently the eigenstructure of M^TM and MM^T. The analysis that follows will show that the eigenstructure depends on the metrics for the inputs and outputs. The metrics are inner product matrices of dimension nxn. Let $\Delta_x = X^TX$ and $\Delta_y = Y^TY$. The eigenvalues and eigenvectors of M^TM can be found by solving:

$$M^TM\phi = \phi\mu$$

If M is a generalized inverse memory, $M^TM = (X^I)^T \; \Delta_y X^I$. Defining ψ through $\phi = X\Psi$ yields the following generalized eigenvalue problem

$$\Delta_Y\Psi = \Delta_X\Psi\mu$$

If M is a correlation memory matrix the inner product is $M^TM = X^T\Delta_Y X$. Defining ξ through $\phi = X^I\xi$ the eigenvalue problem is transformed to the generalized eigenvalue problem

$$\Delta_X\Delta_Y\xi = \xi\mu$$

or $\Delta_Y\xi = \Delta_X^{-1}\xi\mu$ if the inverse exists.

We consider a few special cases then solve the generalized eigenvalue problem for the general case.

Case 1 Orthogonal Inputs $\Delta_X = 1$

In this case, the generalized inverse and correlation matrix memories have the same characteristics. The nonzero eigenvalues of M^TM, μ, are the eigenvalues of Δ_Y.

For high correlation in outputs, some eigenvalues can get large and there will be large noise gain in those directions. The trace of Δ_Y is n so there will be no increase in average output noise strength.

Case 2 Orthogonal Outputs $\Delta_Y = 1$

For high correlation in inputs the μ can get very large in some directions. For the generalized inverse memory the eigenvalues are the eigenvalues of $\Delta_X{}^{-1}$. There will be large noise gains in the directions associated with large μ. The trace of $\Delta_X{}^{-1}$ will be greater than n and there will be a gain in the average noise strength for generalized inverse memories. For the correlation matrix memory the nonzero eigenvalues μ of M^TM are the eigenvalues of Δ_X. The trace of Δ_X is n. So no gain in average output noise will occur for correlation matrix memories.

Case 3 Auto Associative Memory

Here the inputs are the same as the outputs $X_i = Y_i$. The generalized inverse memory M is a projection operator. All of the eigenvalues of M^TM are one and its trace is one. For the correlation matrix memory the eigenvalues of M^TM are the eigenvalues of $(\Delta_X)^2$. If there are inputs which are highly correlated there will be large gains in output noise strength in those directions. The trace of $(\Delta_X)^2$ will be larger than n and a net gain in average output noise strength will result.

Case 4 Heteroassociative But $\Delta_Y = \Delta_X$

This is an interesting special case in which $Y \neq X$; and M need not be square. The generalized inverse memory, M, is not a projection operator but M^TM is a projection operator with unit eigenvalues and a trace equal to n.

For the correlation matrix memory, M^TM has eigenvalues again which are equal to those of $(\Delta_X)^2$. The noise characteristics are identical to those of the autoassociative memory, Case 3, for both the correlation and the generalized inverse memories.

Nonspecial cases require solution of the generalized eigenvalue problem. An analysis and a set of bounds on the generalized eignevalues can be obtained by performing generalized singular value decomposition. For a pair of matrices A and B which have the same number of columns the following decomposition is possible and is called the generalized singular value decomposition of A and B.

$$A = V_A \; aZ^T$$

$$B = V_B \; bZ^T$$

where V_A and V_B are unitary matrices, a and b are matrices whose only nonzero elements lie along the diagonal, and Z is a matrix with linearly independent columns. Without loss of generality, these columns can be assumed to be normalized to unity. For each column of Z, Z_i there is a pair of generalized singular values a_i and b_i where

$$Z_i^T \; A^T \; AZ_i = a_i^2 = Z_i^T \; \Delta_a \; Z_i$$

and

$$Z_i^T \; B^T \; BZ_i = b_i^2 = Z_i \Delta_b Z_i$$

since $V_A^T \; V_A$ and $V_B^T \; V_B$ are unity.

Thus

$$Z_i^T \; \Delta_A \; Z_i = \left(\frac{a_i^2}{b_i^2}\right) Z_i^T \; \Delta_B \; Z_i = \mu_i Z_i^T \; \Delta_B \; Z_i$$

The ratio of the squares of the generalized singular values are the generalized eigenvalues. The generalized singular values are bounded by the minimum and maximum eigenvalues of $A^T B$, α_{min} and α_{max}, and the minimum and maximum eigenvalues of $B^T B$ β_{min} and β_{max},

$$\alpha_{min} \leq a_i^2 \leq \alpha_{max} \text{ and } \beta_{min} \leq b_i^2 \leq \beta_{max}$$

Thus the generalized eigenvalues of $\mu_i = a_i^2/b_i^2$ are bounded by

$$\frac{a_{min}}{b_{max}} \leq \mu_i \leq \frac{a_{max}}{b_{min}}$$

We can apply these results to the generalized inverse associative memory. A generalized singular value decomposition of X and Y yields

$$X = V_x b \ \psi^T$$

$$Y = V_y a \ \psi^T$$

Substituting into $\Delta_Y \psi = \Delta_X \psi \mu$ yields the generalized eigenvalues, $\mu_i = a_i^2/b_i^2$, which are the eigenvalues to the $M^T M$. These are bounded by the ratio of the eigenvalues of Δ_X and Δ_Y, λ_X and λ_Y respectively.

$$\frac{\lambda_Y(\min)}{\lambda_X(\max)} \leq \mu_i \leq \frac{\lambda_Y(\max)}{\lambda_X(\min)}$$

For the correlation matrix memory a generalized singular value decomposition of Y and XI^T is required. Similar analysis yield the following bounds in terms of the eigenvalues of Δ_X and Δ_Y.

$$\lambda_Y(\min) \ \lambda_X(\min) \leq \mu_i \leq \lambda_Y(\max) \ \lambda_X(\max)$$

As these bounds indicate the output noise strength can be much greater or much less that the input noise. In order to achieve the minimum or maximum gain in noise in Generalized Inverse Memories there must be high negative correlation in the output when there is high positive correlation in the inputs and vice versa. For correlation matrix memories, high correlation of same sign in inputs and corresponding outputs will cause large gain in noise. A simple two dimensional example will illustrate the role of the metric matrices in this analysis.

In Figure 1.1 are shown two input vectors X_1 and X_2 and in another plane are two output vectors Y_1 and Y_2.

The metrics Δ_X and Δ_Y are

$$\Delta_X = \begin{pmatrix} 1 & d \\ d & 1 \end{pmatrix}$$

and

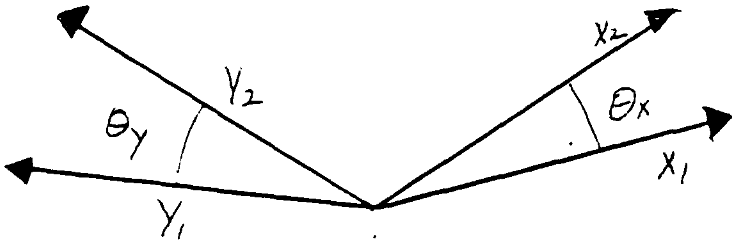
$$\Delta_Y = \begin{pmatrix} 1 & e \\ e & 1 \end{pmatrix}$$

where

$$-1 \leq d \leq 1$$

and

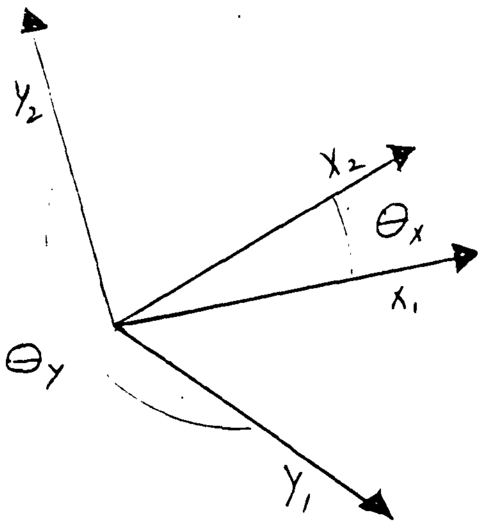
$$-1 \leq e \leq 1$$



$\theta_X = \cos^{-1} d$

$\theta_Y = \cos^{-1} e$

Figure 1.1a. Two Input Vectors and Two Output Vectors that are Positively Correlated d > 0 and F > 0



$\theta_x = \cos^{-1} d$

$\theta_Y = \cos^{-1} e$

Figure 1.1b. Two Input Vectors that are Positive Correlated and Two Output Vectors that are Negatively correlated d > 0 and e < 0

The generalized eigenvalue problem for the case of the pseudo inverse associative memory

$$\Delta_Y \psi_i = \Delta_X \psi_i \mu_i$$

yields eigenvectors

$$\psi_1 = \frac{1}{\sqrt{2}} \binom{1}{1}$$

and

$$\psi_2 = \frac{1}{\sqrt{2}} \binom{1}{-1}$$

with eigenvalues

$$\mu_1 = \left(\frac{1 + e}{1 + d}\right)$$

and

$$\mu_2 = \left(\frac{1 - e}{1 - d}\right)$$

If the inputs are positively correlated and the outputs negatively correlated then μ_2 can become quite large. On the other hand if both inputs and output are positively correlated and $e > d$ then μ_2 is less than 1 and a net reduction in noise will result in the ψ_2 direction. If $d = e$, both eigenvalue are 1 (case 4), and no net gain or reduction occurs.

For the correlation matrix memory

$$\Delta_Y \xi_i = \Delta_X^{-1} \xi_i \mu_1$$

yields eigenvectors

$$\xi_1 = \frac{1}{\sqrt{2}} \binom{1}{1}$$

and

$$\xi_2 = \frac{1}{\sqrt{2}} \binom{1}{-1}$$

with eigenvalues

$$\mu_1 = (1 + e)\ (1 + d)$$

and

$$\mu_2 = (1 - e)(1 - d) \quad .$$

If both inputs and outputs are positively correlated $\mu_1 > 1$, or if both inputs and outputs are negatively correlated $\mu_2 > 1$, and noise gain will occur in the ψ_1 or ψ_2 directions respectively.

The example illustrates how both input and output metrics influence the output noise to input noise ratios. The most dominating factor however is the input metric. In application, the real issue is confusion, will the input plus noise look like an incorrect input? If it does the output will be the incorrect output. The key concern is the effects of noise in directions causing confusion. These directions are those pointing along input differences.

Noise that directly contributed to confusing inputs X_i and X_j lies along $X_i - X_j$, $N_{ij} = n(X_i - X_j)$. The role that the output metric plays is thru the correlation between outputs.

For a generalized inverse memory the confusing output noise is

$$M\ N_{ij} = N^o_{ij} = n(Y_i - Y_j)$$

and thus

$$\frac{\|N_{ij}^{o}\|}{\|N_{i}\|} = \frac{\|Y_i - Y_j\|}{\|X_i - X_j\|}$$

noise in confusing directions increases proportionally to the relative separation of outputs. The example in Figure 1 illustrates this point. When the inputs are positively correlated and the outputs negatively correlated their is a large noise increase in the ψ_2 direction which lies along $M(\bar{X}_1 - \bar{X}_2)$. For this case $(\bar{Y}_1 - \bar{Y}_2)$ being negatively correlated is corresponding large. If both inputs and outputs are positively correlated and the outputs more correlated than the inputs then a net decrease in noise occurs in the $Y_1 - Y_2$ direction, but the distance between Y_1 and Y_2 decreases correspondingly.

This analysis suggests the real concerns with noise are the magnitudes of the input noise in confusing directions and the distance between two inputs. If the magnitude of the noise in a confusing direction exceeds the magnitude of the distance between the two input vectors, confusion will occur. Otherwise it will not. The analysis for correlation matrix memories is complicated by the cross talk. For input noise pointing along a confusing direction, N_{ij} the output noise no longer lies exactly along the confusing direction in the output space.

$$\frac{\|N_{ij}^{o}\|}{\|N_{ij}\|} = \frac{\|(Y_i - Y_j) + (C_i - C_j)\|}{\|X_i - X_j\|}$$

The difference in cross talk can be either positive or negative and the component of this difference along $Y_1 - Y_2$ can also be either positive or negative.

Reference for Section 1

1. Adi Ben-Israel and Thomas N.E. Greville, "Generalized Inverses: Theory and of Applications," John Wiley (1974).

2. T. Kohonen, "Self Organization and Associative Memory," Springer-Verlag, (1984).

3. G.S. Stiles and Dong-L.H. Denq, "On The Effect of Noise and Modle Penrose Generalized Inverse Associative Memory," IEEE PAMI 7, 358 (1985).

2. ADAPTIVE ASSOCIATIVE MEMORY MODELS WITH POTENTIAL OPTICAL IMPLEMENTATION

A pseudo inverse matrix memory model has been successfully implemented optically at NRL.[3] The algorithm used there is adaptive. Recall is a single matrix multiplication. Let the input vectors, X, have ℓ components and the output vectors Y have p components then the p x ℓ memory matrix, M, recalls Y by matrix multiplying X

$$Y = MX$$

During the learning phase of operation M can be found from a solution to $X^T M^T = Y^T$. This is a standard linear estimation problem. However if the number of input output pairs is less than the number of components in the input vectors, the system is undetermined and M is not uniquely determined. A unique solution that gives the memory matrix M minimum Frobenius norm $\|M\|_F$ is obtained from the Moore Penrose pseudo inverse of X, X^I.

This memory can be found through an iterative algorithm which uses one row of X^T and Y^T at a time, i.e., it uses only one input output pair. The algorithm was first introduced by Kacrmarz[1] and later used first in associative memories by Widrow.[2]

The key step in the algorithm is

$$M^T (k + 1) = M^T (k) + \lambda X_1 (Y_k^T - X_k^T M_k^T)$$

This algorithm is currently implemented at NRL.[3] With minor modifications the algorithm can be used to find the pseudo inverse of X or to be a novelty filter for X. By choosing Y to be the n x n unit matrix the algorithm will yield the pseudo inverse of X. By initialized M(o) to the $\ell \times \ell$ unit matrix and setting Y to zero M becomes a novelty filter. The method of Kaczmarz has spawned a large variety of algorithms for large systems and their applications. These methods are known as row action methods[4-7] since in a single step only one row of the matrics are required. Some basic algorithms that may have use in implementation are included in the following section.

Row Action Methods

We outline here a few of the row action methods which may be applicable to optical implementation of associative memory and recall. Since the algorithms can be used for both purposes we describe them here in a generic notation commonly used for systems of linear equations

$$A\,x = b$$

This system may be underdetermined, greatly overdetermined with possibility of inconsistencies (self contradiction) and it can be ill-conditioned. There may be reason to doubt that the exact algebraic solution is the desired one, especially if the data contains measurement inaccuracies and or noise, distortion, and missing data. These other approaches include;

(i) Constrained Minimization

$$\mathrm{Min}\ \|Ax - b\|$$

subject to $x \in S$. By $x \in S$, we mean it may be required to satisfy some set of constraints for example box constraints.

$$t_i \leq x_i \leq p_i$$

(ii) Regularization

Minimize

$$\left[f\ (x) + \lambda\ \|Ax - B\|\right]$$

where f (x) is usually the square norm of x, $\|x\|^2$, and $x \varepsilon S$

(iii) Feasibility

For each row vector a_i^T of A

$$\left(b_i - \gamma_i\right) \leq a_i^T x \leq \left(b_i + \xi_i\right) \qquad x \varepsilon S$$

The γ_i and ε_i are picked so that the feasibility region is not empty.

(iv) Optimization

Minimize F(x) subject to

$$\left(b_i - \gamma_i\right) \leq a_i^T x \leq \left(b_i - \xi_i\right) \qquad x \varepsilon S$$

The basic algorithm is the Kaczmarz algorithm. It solves Ax = b.

In this notation the n^{th} step is

$$x^{n+1} = x^n + \mu_k\ a_k\left(b_k - a_k{}^T x^n\right)$$

where $\mu_k = \lambda_k / \|a_k\|^2$ and λ_k called the relaxation parameter is bounded by $0 < \lambda_k < 2$. In the field of Image reconstruction from projections this algorithm is known as ART.[8]

The most similar algorithm is the relaxation method of Agmon[9] and Motskin and Sihoenberg[10] for the linear feasibility problem. Collecting all constraints and expressing them as

$$Ax \leq b$$

the basic step is

$$x^{n+1} = x^{n} + a_k^T G_k$$

where

$$g^k = \lambda_k(b_k - a_k^T x^n)$$

for the right hand side negative and

$$g^k = 0$$

otherwise.

The interval feasibility problem can be implemented by repeating the rows, one positive to satisfy $a_i^T x \leq b_i + \xi_i$ and one negative to satisfy $-a_i^T x \leq (b_i - \gamma_i)$.

A method for regularization has been developed by Herman et al.,[11] The procedure minimizes $f(x) = \|x\|^2 + \gamma^2 \|Ax - b\|^2$ the n^{th} step is

$$z^{n+1} = z^{n} + e_k g_k$$

$$x^{n+1} = x^{n} + \gamma a_k^T g_k$$

where e_k is the vector with elements zero except for the kth which is unity, i.e., the kth column of the unit matrix. z^0 is arbitrary and $x^0 = \gamma a^T z^0$. g_k is found from

$$g_k = \mu_k + [\gamma (b_k - a_k)^T x^n] - z^n$$

where

$$\mu_k = \frac{\lambda_k}{1 + \gamma \| a_k \|^2}$$

Each of these methods may be implemented optically. The feasibility method however requires a thresholding step for g_k. The regularization method may require a second spatial light modulator to store z.

All of the above algorithms can be made nonlinear by requiring x to satisfy constraints $x \varepsilon S$. To implement the box constraints for example the output at the end of each step, x^n, can the threshold so that

$$\hat{x}_i^n = x_i^n \qquad \text{if} \qquad t_i \leq x_i^n \leq p_i$$

$$\hat{x}_i^n = p_i \qquad \text{if} \qquad x_i^n \leq p_i$$

$$\hat{x}_i^n = t_i \qquad \text{if} \qquad x_i^n \leq t_i$$

and $\hat{x}^n$ is used to compute x^{n+1}. Simple thresholding can be used to solve these problems subject to the constraint conditions.

References for Section 2

1. S. Kaczmarz, Bull. Acad. Polon Sci. Lett. A, 35, 355 (1937).

2. B. Widrow, Self-Organizing Systems, M.C. Yarrts, et al., Eds. Spartan Books, Washington, 435, (1962).

3. A.D. Fisher and C.L. Giles. "Optical Adaptive Associative Computer Architecture." Proc. IEEE Compcon Meeting, Feb. 1985.

4. Yair Censor, Finite Series Expansion Reconstruction Method, Proc. IEEE, 71, 409 (1983).

5. Arnold Lent and Yair Censor. "Extenstions of Hildreth's Row Action Method for Quadratic Programming", SIAM J. Control and Optimization, 18, 444 (1980).

6. Ronald Schafer, Russell Mersereau and Mark Richards. Constrained Iterative Restoration Algorithms, Proc. IEEE 69, 432 (1981).

7. Tommy Elfving. On Some Methods of Entropy Maximization and Matrix Scaling Lin. Algebra and its Applications, 34, 321 (1980).

8. R. Gordon P. Bender and G.T. Herman, "Algebraic Reconstruction Techniques (ART) for Three Dimensional Electron Microscopy and X-ray Photography." J. Theoret, Biol. 29, 471 (1970).

9. S. Agmon, "The Relaxation Method for Linear Inequalities." Canad, J. Math 6, 382 (1954).

10. T.S. Molzkin and I.J. Schoenberg, "The Relaxation Method for Linear Inequalities," Canad. J. Math 6,, 393 (1954).

3. ON OPTIMAL ASSOCIATIVE RECALL FROM AN INCOMPLETE INPUT VECTOR

An issue of concern in associative memory design is its robustness against noisy input data. Without a noise model the problem is quite formidable. A problem that is somewhat less difficult, that of recall from input patterns that have been masked so that some components are missing (i.e., equal zero), has been attacked more frequently. An interesting paper by Marakami et. al., found an optimal associative memory for recall of a pattern from a input pattern that had a known fixed set of components equal to zero. The memory is optimal is the sense of best linear unbiased estimate (i.e., least squares). If we let the input vectors have a total of n components and let n - s of the components be masked, the optimal associative memory for inputs with only s out of n components present is

$$^{s}M = (n - 1)\ YX^{T}\ [(s - 1)\ XX^{T} + (n - s)\ R]^{I}$$

where R is the diagonal of XX^T

This result can be simplified if the rows of X are equilibrated so that they have unit length. If we assume that this has been done then R = 1 and

$$^{s}M = (n - 1)\ YX^{T}\ [(s - 1)\ (X^{T}X) + (n - s)\ 1]^{I}\ X^{T}$$

This expression is valid for $0 < s \leq n$. For s = 1, only 1 component unmarked the memory reduces to a correlation matrix memory. For s = n it reduces to the Moore Penrose pseudo inverse memory. For $s \neq n$ the memory resembles a regularization inverse. In fact this memory can be arrived at from a constrained minimization problem. Minimize the average noise output strength of the memory, keeping the total mean square error in memory components equal to R. The problem takes the form

$$\mathrm{Min}\ \|M^T M\|_F$$

subject to

$$\|MX - Y\|_F = R$$

where

$$\|\cdot\|_F$$

is the Frobenius Norm.

When used in recall the memory also solves a minimization problem. We can factor the solution into two steps. We write ^sM as

$$^s_M = \gamma^{-1}\ Y\ [X^T X + \propto I]^I X^T$$

where

$$\gamma = (\frac{s - 1}{n - 1})$$

and

$$\propto = (\frac{n - s}{s - 1})$$

Step (1) is solve

$$(X^TX + \alpha I)^I X^T\hat{X} = \hat{t} \quad .$$

then step (2) multiply $\hat{t}$ by Y to get the output $\hat{Y}$.

Step (1) is equivalent to solving

$$\hat{X} = Xt$$

using regularization techniques, that is it is equivalent to

$$\text{Min } \|t\|$$

subject to $\|Xt - \hat{X}\| = R$

By keeping the elements of t small we keep the memory outputs smooth, we reduce the likelihood of large oscillatory mixing of output components. α is the reciprocal of the Lagrange multiplier which makes the solution satisfy the constraints. In applications, one typically solves step 1 for a set of α's and then settles on a residual and norm which seems reasonable. The optimum associative memory selects α to be large compared to typical regularization values and hence excepts a rather large residual for minimum noise output. Given that n - s components of the input are known to be wrong this is quite reasonable. An algorithm which would allow optical implementation of this memory is given in Section 2.

References for Section 3

1. K. Murakami, S. Akaishi, and T. Aibara, "On Optimal Asociative Recall By An Incomplete Key," Biol. Cybenetics <u>30</u>, 95 (1978).

4. PROJECTION OPERATORS AND SUBSPACE METHODS OF PATTERN RECOGNITION

A standard classification problem is that of identifying an unknown input as being a member of a previously learned class. This identification process is typically hampered by corrupting noise as well as the event that a current unknown input may be either totally new and not a member of any of the previously learned classes, or may be a mixed signal which belongs to two or more of the previously learned classes.

The basic concept of similarity among patterns is that the features of similar patterns are similar. If these features are used to form a feature space, each pattern can be represented by a vector in this space. Similar patterns will have similar vectors, that is, vectors which are all in some local region of the space and are close together in terms of some metric. In subspace methods the feature space is divided into subspaces, one subspace for each class of patterns. Ideally these subspaces would have no intersections. The subspace would be mutually exclusive pattern vectors of a class would lie totally within the subspace of that class. The distance of an input pattern from a class subspace is a measure of its similarity to the patterns of the class.

Orthogonal projection methods,[1-9] which use a different orthogonal projector to represent the subspace spanned by a class have been developed and extensively studied. For potential mixed patterns, the lengths of the projections of an unknown onto each of the class subspaces can be related to the amounts of the individual class that are present in the unknown. The relation is not a simple one when the classes are not orthogonal to one another. The use of oblique projectors rather than the more common orthogonal projectors yields an attractive alternative. The major difficulty with orthogonal projection is that unless different class subspaces are orthogonal, there can be a large projection of an unknown of one class onto the subspace of another class. This is a problem for both inputs which are mixtures and pure inputs which are noisy. We show in Subsection 4.4 that for more than two classes, noisy inputs can cause misclassification in orthogonal projections while the same inputs are correctly classified by the corresponding set of oblique projections. For mixtures there is another complication due to lack of orthogonality of subspaces. A possible identifier for a mixture of two classes would be the length of the projection of the unknown onto the subspace spanned by the union of the two class subspaces. However, for nonorthogonal subspaces the projection onto the union is not the sum of the projections onto each class.

The oblique projection techniques are robust against noise and will correctly predict the input until the noise has a larger component in another class subspace. Orthogonal projection can fail before the noise is this large because of cross talk between classes.

The major advantage of oblique projection operators is that the feature space can be divided into independent mutually exclusive subspaces by a set of mutually exclusive oblique projection operators. These operators will give zero projection for a pattern belonging to a different subspace or class.

The use of oblique projectors in pattern recognition has been somewhat limited to date[10] although the mathematical theory has been developed.[11-14]

To introduce the concepts of oblique projections the elementary properties of idempotent operators are reviewed. The relation of these projections to least square techniques and regression is illustrated. A constructive scheme for the projections based on least squares analysis presented is in Subsection 4.3.

In Subsection 4.4 a geometric interpretation of oblique projections is presented and a comparison between orthogonal and oblique projection classification schemes. A general method of construction of oblique projection operators is presented in Subsection 4.5. The subspace method can be generalized to include bounded regions or zones by the introduction of constrained projections. These add nonlinearity and provide a means of adding additional information about a class when it is known. The implementation is discussed in Subsection 4.6. The relation between generalized inverses and projections is discussed in Subsection 4.7. This introduction of weighted features into subspace methods is discussed in Subsection 4.8.

4.1 Elementary Properties of Idempotent Operators and Projectors

The general properties of idempotent operators and the decomposition of spaces into linearly independent subspaces is reviewed. Such operators play a natural role in pattern recognition.

An operator P is idempotent if $P^2 = P$. If P is idempotent then its complement I - P is also idempotent.

$$(I - P)\ (I - P) = I - 2P + P^2 = I - 2P + P = I - P \quad .$$

The major feature of idempotency and the use of idempotent operators in pattern recognition stems from the following.

An idempotent operator P decomposes a space, ϕ, into two linearly independent subspaces, ϕ_1 and ϕ_2 such that if V belongs to ϕ_1 then PV = V and if V belongs to ϕ_2, PV = 0. Every vector y in ϕ can be decomposed into $y = y_1 + y_2$ where

$$Py = y_1 \quad ,$$

and

$$(I - P)\ y = y_2 \quad .$$

y_1 belongs to ϕ_1 and y_2 belongs to ϕ_2. It is useful to introduce a votation change before generalizing to several subspaces. Let $P_1 = P$ and $P_2 = 1-P$, then $P_1 y = y_1$ and $P_2 y = y_2$.

The desired properties for an independent subspace method for pattern recognition are provided by a set of mutually exclusive idempotent operators.

Given a set of K idempotent operators P_1, P_2, ... P_K such that

$$P_I P_J = \delta_{IJ} P_J$$

where

δ_{ij} is the Kronecker delta

and

$$\sum_{I}^{K} P_I = P_\phi$$

then ϕ decomposes into k linearly independent subspaces ϕ_1 ϕ_2 ... ϕ_K. P_ϕ is the total projector. It projects a function into the space ϕ. If ϕ is the entire space $P_\phi = I$, the identity operator.

Each of the K subspaces are defined as the set of all vectors (patterns) belonging to the particular ϕ_i. The subspaces are independent but not required or assumed to be orthogonal in order for the mutual exclusive property to hold. If the pattern V belongs to subspace ϕ_J (class J) then

$$P_J V = V$$

and

$$P_I V = 0$$

for I ≠ J.

For such a classification scheme, patterns that are mixtures of two or more classes can be identified. Let

$$V = V_L + V_M$$

then

$$P_L V = V_L \quad ; \quad P_M V = V_M$$

and

$$P_I V = 0$$

for I ≠ L or I ≠ M.

The key to the use of operator techniques is the determination of the desired operators and finding representations for them. If the subspace ϕ_K are orthogonal the idempotent operators will be symmetric. These operators are called orthogonal projection operators. If the subspace ϕ_K are not orthogonal, the idempotent operators will not be symmetric. These idempotent operators are called oblique projection operators.

Characteristics and construction methods for oblique and orthogonal projection operators is further developed below.

4.2 Orthogonal Subspaces and Orthogonal Projection Operators

Two subspaces are orthogonal if every vector belonging to subspace ϕ_1 is orthogonal to every vector belonging to subspace ϕ_2. Orthogonal subspaces are special cases of linearly independent subspaces, and hence can be analyzed using idempotent operators. The idempotent operators in turn will have special characteristics. Let R^n be divided into two orthogonal subspaces. ϕ_1 and ϕ_2. There exist idempotent operators P_1 and $P_2 = I - P_1$ such that

$$P_1U = U_1$$

$$P_2U = U_2$$

and

$$U_1^T U_2 = 0 \quad .$$

The special characteristic is that the idempotent operators associated with orthogonal subspaces must be symmetric.

To show this we consider two vectors U and V and their decompositions

$$U = U_1 + U_2 = P_1U + P_2U$$

and

$$V = W_1 + W_2 = P_1V + P_2V \quad .$$

Due to the orthogonality of the subspaces

$$U_1^T V = U_1^T V_1 = U^T V_1$$

as

$$U_1^T V_2 = 0$$

and

$$U_2^T V_1 = 0$$

The scalar product $U_1^T V$ can be written as

$$(P_1U)^T V = U^T P_1{}^T V \quad .$$

The scalar product U^TV_1 can be written as

$$U^T P_1 V \quad .$$

Since the two scalar products are equal

$$P_1^T = P_1 \quad ,$$

that is P_1 is symmetric. Operators which are both idempotent and symmetric are called orthogonal projection operators. The orthogonal projection of a vector onto a subspace has a simple geometric interpretation. For $P_1V = V_1$, $V - V_1$ is clearly orthogonal to V. The key property is that the orthogonal projection, V_1, is that vector of ϕ_1 which is closest in distance to V. $(V - V_1)$ is the shortest vector between V and the space ϕ_1. For any nonorthogonal (oblique) projection, $\hat{V}_1$, $(V - \hat{V}_1)$ is longer than $(V - V_1)$.

4.3 Projection Operators and Linear Estimation

Let ϕ be a subspace of R^n and let $\{W_\phi\}$ be a basis for ϕ. Then the orthogonal projection of a function f onto ϕ can be obtained through the least squares procedure

$$f = WC + R$$

where the expansion coefficients are given by

$$C = (W^TW)^{-1} W^Tf = \Delta^{-1} W^Tf \quad .$$

The desired projection operator is

$$P_\phi = W \Delta^{-1} W^T \quad .$$

Note

$$P = P^T$$

and

$$P^2 = W \Delta^{-1} \left(W^T U\right) \Delta^{-1} W = P \quad .$$

Thus we have

$$\hat{y} = Py \text{ and } R = \left(1 - P\right)y \quad .$$

The K dimensional subspace ϕ may be broken down into k one dimensional subspaces associ: ted with each W_K. If these subspaces are orthogonal then Δ is the unit matrix and

$$P_\phi = \sum_{I=1}^{K} W_I W_I^T = \sum_{I}^{K} P_I \quad .$$

That is if the one dimensional subspaces are orthogonal $P\phi$ can be written as the sum of K one dimensional orthogonal projection operators.

It is always possible to generate orthogonal subspaces by using standard orthogonalization procedures, Gam Schmidt for example. However, if the nonorthogonal basis has a model interpretation associated with it, this interpretation may be lost on orthogonalization.

When the subspaces are not orthogonal, the projection onto the subspaces ϕ_I will not be orthogonal. The individual projections will be oblique. The metric Δ and its inverse Δ^{-1} can be partitioned such that the projection operatore P_ϕ can be expressed as a sum of oblique projection operators

$$P_\phi = \sum_{I} O_I$$

where

$$O_I = W_I \, \Delta_{II}{}^{-1} \, W_I^T + \sum_{J \neq I} W_I \, \Delta_{IJ}{}^{-1} \, W_J^T \quad .$$

The operatore O_I is the oblique projection operator for subspace ϕ_I.

$$O_I^2 = O_I$$

but

$$O_I^T \neq O_I \quad .$$

The oblique projection operators have the following properties

$$O_I \, W_J = \delta_{IJ} W_I \tag{a}$$

$$\left(1 - O_I\right) W_I = 0 \tag{b}$$

$$O_I \, O_J = \delta_{IJ} O_I \tag{c}$$

$$P_\Phi = \sum_I O_I \tag{d}$$

Property a is key to the use of these operations for analysis of mixed unknowns, unknowns with membership in more than one class. If unknown

$$U = W_I \, C_I + W_J \, C_J$$

then

$$O_I U = W_I \, C_I$$

$$O_J U = W_J C_J$$

and

$$O_K U = 0$$

for all

$$K \neq I \text{ or } J \quad .$$

It is useful to compare this result with the prediction of independent orthogonal projections on a mixed unknown. The orthogonal projection operation for subspace ϕ_1 is

$$P_1 = W_1 W_1^T \quad .$$

The projection of unknown U onto ϕ_1 is

$$P_1 U = W_1 (C_2 + \Delta_{12} C_2)$$

Since, in general,

$$P_I W_J = W_I \Delta_{IJ}$$

where

$$\Delta_{IJ} = W_I^T W_J$$

is the overlap between the subspaces.

Thus

$$P_1 U = W_1 B_1$$

where

$$B_1 \neq C_1 \quad .$$

The effect of orthogonal projection is to include some of the ϕ_2 component into the projection. This is an undesirable property. The oblique projection approach is more appropriate for nonorthogonal subspaces.

We can conclude that the best linear unbiased approximation of a function f by a basic set of functions $\{W_I\}$ is the orthogonal projection of f onto the subspace, ϕ, by the basic $\{W_I\}$. This orthogonal projector, P, is the sum of individual projectors, one for each of the component subspaces of ϕ. The individual component projectors will be orthogonal if the basis is orthogonal. The component projections will be oblique if the basis is nonorthogonal.

4.4 Geometric Interpretation of Oblique Projections

Geometric interpretations of oblique and orthogonal projection operators are presented. The relationship between coefficients obtained for orthogonal and oblique projections as compared for simple R^2 and R^3 examples. We have shown that the minimum distance classifiers defined in terms of oblique and orthogonal projection methods agree only for the two class case in R^2. A simple example in R^3 shows that orthogonal projections will not predict correct classifications for nonorthogonal classes.

Oblique projections require two subspaces for their definition, the subspace into which projection occurs and the subspace into which the adjoint projection occurs. We label these subspaces ϕ and Ψ, respectively.

Of is the projection of f onto ϕ which is along $\Psi^{\perp}$. $O^T f$ is the projection of f onto Ψ which is along $\phi^{\perp}$. For a function f and the oblique projection operator O, Of lies along ϕ. See Figure 4.1. The projection is along, $\Psi^{\perp}$, the space orthogonal to Ψ. $\Psi^{\perp}$ is the orthogonal complement of Ψ. Ot is also the oblique complement of ϕ. The adjoint operator, O^T, projects f onto Ψ. This projection is along or parallel to, $\phi^{\perp}$, the orthogonal complement of ϕ. $\phi^{\perp}$ is also the oblique complement of Ψ. If ϕ and Ψ are equal, then the projection is orthogonal. One idempotent operator generates two pairs of complementary projections $(O, 1-O)$ and $(O^T, 1-O^T)$

which projects onto complementary subspace pairs $(\phi, 1-\phi)$ and $(\Psi, 1-\Psi)$ respectively.

We concentrate now on a complimentary pair and their relation to orthogonal projections. Consider the projections O_I and $O_J = 1 - O_I$ on the independent but nonorthogonal subspaces ϕ_I and $\phi_J = (1-\phi_I)$. In Figure 4.2 an R^2 example is given. Both oblique projections of a function f onto ϕ_I and ϕ_J are illustrated as well as independent orthogonal projections P_If and P_Jf. We illustrate that for the two class example in R^2 if $C_I > C_J$ then $B_I > B_J$. This means that either the orthogonal projector or the oblique projector will lead to the same prediction if used as

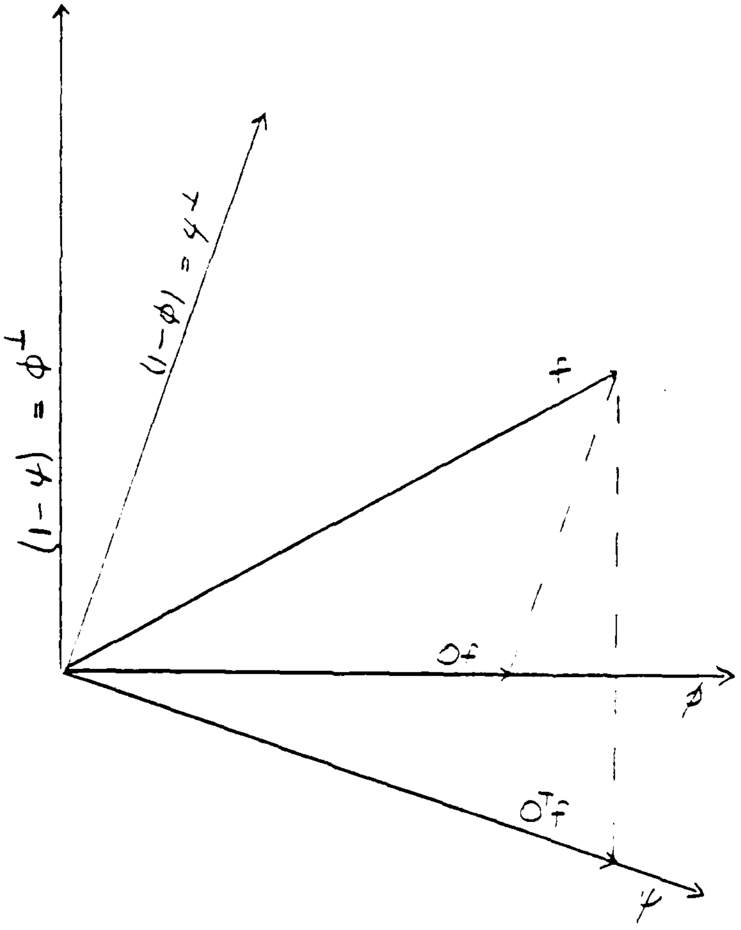


Figure 4.1. Oblique Projections

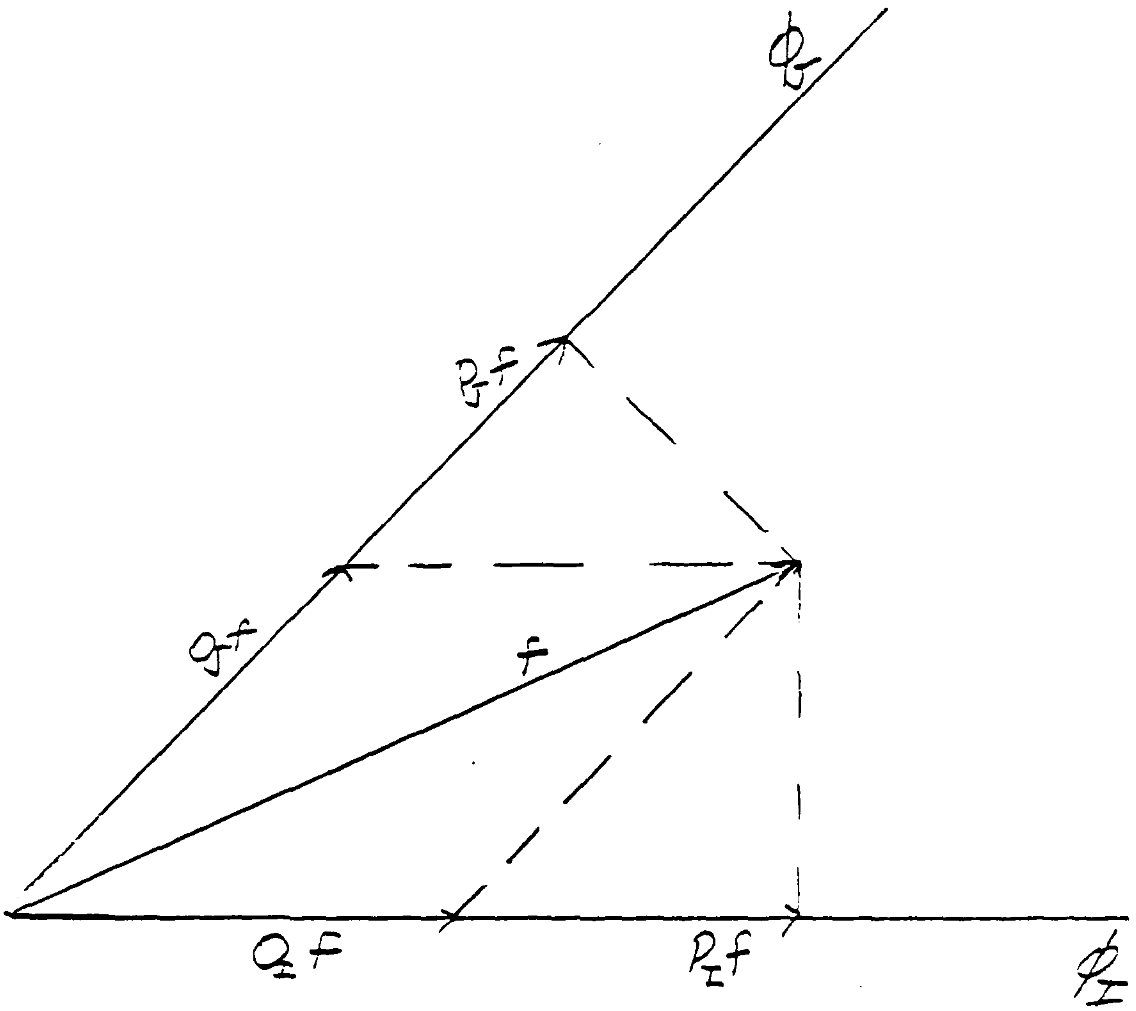


Figure 4.2. Oblique and Orthogonal Projections of f onto the Nonorthogonal Complimentary Subspaces Φ_I and Φ_J

a minimum distance classifier for a two class problem. We will also show that this does not generalize to a three class problem. If we let W_I be a basis for ϕ_I and W_J be a basis for ϕ_J then

$$O_I f = W_I C_I$$

$$O_J f = W_J C_J$$

$$P_I f = W_I B_I$$

and

$$P_J f = W_J B_J$$

The projections coefficient B and C are related through the metric

$$B = \Delta C$$

For Δ equal to $\begin{pmatrix} 1 & d \\ d & 1 \end{pmatrix}$

where the W vectors are assumed to be normal vectors and d is the overlap between classes $0 < d < 1$.

$$B_I = C_I + d\ C_J$$

$$B_J = C_J + d\ C_I$$

the difference $B_I - B_J$ is proportional to $C_I - C_J$. If class selection is based on the larger projection, then for $C_I > C_J$; $B_I > B_J$ and both oblique and orthogonal projection predict class I. These results do not extend to the three class problem. Consider the metric Δ

$$\Delta = \begin{pmatrix} 1 & d & e \\ d & 1 & f \\ e & f & 1 \end{pmatrix}$$

then

$$B_I = C_I + dC_J + eC_K$$

$$B_J = C_J + dC_I + fC_K$$

and

$$B_I - B_J = (C_I - C_J)(1-d) + (e-f)C_K$$

The coefficients $B_I = w_I^T f$ are unnormalized correlation coefficients. The correlation between W_I and the input f is

$$r_I = \frac{B_I}{\|W\|_I \, \|f\|} \quad .$$

The magnitude of B_I is the length of the orthogonal projection of f into the subspace ϕ_I. The coefficients C_I are related to signal strengths. If both W_I and f are normalized to one the square of the coefficient $|C_I|^2$ is the best unbiased estimate of the strength of W_I in the function f. If the functions f are mixtures of subspaces, then the deconvolution of these into subspace components requires the oblique projectors. The strength of the subspace contributions is given by $|C_I|^2$.

The fact that the correlations B_I do not necessarily follow the C_I in relative magnitude raises concerns about the use of orthogonal projection as a classification technique when inputs unknowns are noisy. If a pure input from subspace ϕ_I is corrupted by noise which has components in subspaces ϕ_J and ϕ_K, the relative magnitude of the correlation coefficients may not follow the relative strengths of the signal and noise components.

4.5 General Method of Construction of Oblique Projectors

A general method for generation of oblique projections is presented. This approach is independent of a least square model. An oblique projection operator can be defined in term of a pair of subspaces ϕ and Ψ of equal dimension. Given W as a basis for ϕ and Z as a basis for Ψ the projector O

$$O = W(Z^T W)^I \ Z^T$$

is defined where $(Z^T W)^I$ is a generalized inverse of $(Z^T W)$. O is idempotent and therefore a projector.

$$O^2 = W(Z^T W)^I Z^T \ (Z^T W)^I \ Z^T$$

$$= W(Z^T W)^I \ Z^T = O$$

In applications the subspaces ϕ_A and ϕ_B of dimension n_A and n_B respectively will be known. The basis W_A for ϕ_A and W_B for ϕ_B can be determined. To construct the desired projection operator O_A for which

$$O_A W_A = W_A$$

and

$$O_A W_B = 0$$

A basis Z_A is required. This basis will span the n_A dimensional subspace Ψ_A.

The subspace Ψ_A must have an orthogonal projection into ϕ_A and must be orthogonal to ϕ_B. The desired basis Z_A can be constructed by taking the basis W_A and orthogonalizing it to the basis W_B,

$$Z_A = W_A - W_B Q_A$$

where

$$Q = (W_B^T W_B)^{-1} \ (W_B^T W_A) = (\Delta_{BB})^{-1} \ \Delta_{BA}$$

is the desired

$$n_B \times n_A$$

orthogonalization matrix.

Similarly

$$Z_B = W_B - W_A Q_B$$

where

$$Q_B = (W_A^T W_A)^{-1} \ (W_A^T W_B)$$

After a bit of algebraic manipulation it can be shown that 0 is the orthogonal projector for the space $\Phi = \phi_A + \phi_B$. This generation method is independent of initial basis selection for ϕ_A and ϕ_B and is independent of the orthogonalization method for generation of Z_A and Z_B. This is the consequence of the invariance of 0 to nonsingular transformation of the bases

W and Z. Let X and Y be nonsingular. Then let
$\hat{W} = WX$ and $\hat{Z} = ZY$ then

$$\hat{O} = \hat{W}\ (\hat{Z}^T\hat{W})^I\ \hat{Z}^T$$

$$= WX\ (Y^T Z^T WX)^I\ YZ^T$$

$$= WXX^{-1}\ (Z^T W) Y^{-1}\ YZ^T = O$$

Therefore $O = O$ and projection is independent of the bases W and Z.

4.6 Constrained Projection Operators

The representation of a class as a subspace of its features does not always contain all of the information that in known about the class. For example with spectral classes, the absorbances are always positive, but the subspaces spanned by a set of spectra include both positive and negative absorbances. In order to avoid nonphysical spectral estimates as well as incorrect classifications, it is useful to constrain the absorbances to be positive. We develop below the projection method for two types of constraints, inequality constraints and equality constraints. The constraint methods are developed in the framework of constrained least squares. The result is a set of operators for constrained oblique projection.

The solution to the least square problem can be expressed as

$$\text{minimize } (b-Ax)^T\ (b-Ax)$$

with respect to x. The solution $A\hat{x}$ is given by

$$A\hat{x} = A(A^TA)^{-1}A^Tb = Pb$$

where P is the orthogonal projector. The inclusion of constraints can be accomplished using Lagrangian techniques.[15] Constraints of the form $Gx = 0$, equality constraints and $Gx \geq 0$ inequality constraints will be considered. The more general constraints relations $Gx \geq h$ can be converted to the above by the transformation $x = y + G^Ih$ where G^I is the pseudoinverse of G.

Equality Constraints

The least square method with equality constraints can be expressed as

$$\text{minimize } (b-Ax)^T\ (b-Ax)$$

subject to $Gx = 0$.

The associated Lagrangian is

$$L(x,\lambda) = \frac{1}{2}\, x^TA^TAx - x^TA^Tb - \lambda Gx$$

A saddlepoint solution of $L(x,\lambda)$ is obtained if

$$A^TAx + A^Tb - G^T\lambda = 0$$

$$Gx = 0$$

and

$$\lambda \geq 0$$

λ is the vector of Lagrange Multipliers.

The solution x of the constrained problem can be given in terms of the unconstrained problem as

$$x = \hat{x} + (A^T A)^{-1} G^T \lambda$$

where λ is given by

$$\lambda = - [G(A^T A)^{-1} G^T]^{-1} \, G\hat{x}$$

The solution Ax can be expressed as

$$Ax = A \; Q(A^T A)^{-1} Ab = P_Q b$$

where

$$Q = 1 - (A^T A)^{-1} G^T [G(A^T A)^{-1} G^T]^{-1} G$$

A bit of algebra will reveal that Q is idempotent and hence a projector.

$$Q^2 = Q \text{ and further } P_Q^2 = P_Q$$

Q projects the unrestricted solution $\hat{x}$ onto the constraint subspace.

The residual includes the unrestricted residual as well as the residual arising from projection onto the

$$R_Q = A(1-Q)(A^T A)^{-1} A^T b$$

subspace that violates the constraints.

Inequality Constraints

For inequality constraints the optimization problem is

$$\text{Minimize } (b-Ax)^T(b-Ax)$$

$$\text{Subject to } Gx \geq 0$$

Following the method outlined for equality constraints the solution for x and λ are obtained from the solution of

$$(A^TA)X - A^Tb - G^T\lambda = 0$$

$$Gx \geq 0$$

$$\lambda Gx = 0$$

and

$$\lambda \geq 0$$

The major difference between equality and inequality constraints is that the inequality constraints need not be active. If $\hat{x}$ satisfies the constraints in that

$$G\hat{x} > 0$$

then

$$x = \hat{x}$$

$\lambda = 0$ is the solution. In general for every component $\hat{x}_i$ of $\hat{x}$ that is greater than zero, the corresponding Lagrange multiplier λ_i is zero. The nonzero λ_i corresponds to active constraints $x_i = 0$. These correspond to the unrestricted solutions $\hat{x}_i$ violating the constraints, $\hat{G}_i x_i \leq 0$. The nonzero λ_i are given as before for equality constraints and x, λ and P_Q have the same properties. Q however depends on b, if for example

$$G\hat{x} = G(A^TA)^{-1}A^Tb > 0$$

then

$$Q = 1 \quad .$$

Without loss of generality we can assume that the first k constraints are inactive and the remaining constraints are active. Then G can be partitioned into $[G_1G_2]^T$ where G_2 corresponds to the active constraints and the nonzero Lagrangian components solved for from

$$\bar{\lambda} = - [G_2(A^TA)^{-1} G_2^T]^{-1} G_2^T \hat{x}$$

The projection operates P_Q for the constrained least square problem can be considered as a sum of oblique projection operators, one for each class as in the case of unconstrained least squares.

4.7 Relation Between Generalized Inverses and Projection Methods

A relationship between certain generalized inverses and constrained oblique projection operators was identified. The variety of generalized inverses suggests that a large variety of potentially useful projection operators can be generated.

Oblique projectors and constraints can be considered in terms of generalized inverses of matrices. A projection operator can be written as $P_W = W W^I$ where W^I is the generalized inverse of W. The selection of the Moore Penrose inverse

$$W^I = (W^T W)^I W^T = (W^T W)^{-1} W^T = \Delta^{-1} W^T$$

leads to the orthogonal projector $P_W = W\Delta^{-1}W^T$. The Moore Penrose inverse is identical to the true inverse for nonsingular matrices. The existence of Δ^{-1} is based on the linear independence of the columns of w. The Moore Penrose inverse X satisfies the following four properties:

$$AXA = A \tag{1}$$

$$XAX = X \tag{2}$$

$$(AX)^T = AX \tag{3}$$

$$(XA)^T = XA \tag{4}$$

An example of a generalized inverse that does not satisfy the four conditions is

$$A^I = (Z^T A)^I Z^T$$

This inverse satisfies conditions 1, 2 and 4 but not 3 as

$$0 = A(Z^T A)^I Z^T \neq 0^T = Z(A^T Z)^I A^T$$

Condition (3) is equivalent to requiring $A A^I$ to be an orthogonal projection. The oblique projections themselves contain a generalized inverse. The inverse $A^I = (Z^T A)^I Z^T$ is called a 1, 2, 4 inverse.[14] In this nomenclature the Moore Penrose inverse is a 1, 2, 3, 4 inverse.

The construction of Z as

$$Z = A - B(B^T B)^{-1} (B^T A)$$

insures that $Z^T A$ will be singular only if the subspaces spanned by A and B are not independent. If the subspaces are dependent then one or more of the columns of Z will be of zero length. For nonsingular $Z^T A$, the generalized inverse $(Z^T A)^I$ can be replaced with the true inverse in A^I and 0 respectively. The Moore Penrose inverse is identical to the true inverse in this case. If we let $(Z^T A)^I = (Y^T Z^T A)^I Y^T$ and substitute this inverse into 0 we have

$$0 = A(Z^T A)^I Z^T = A(Y^T Z^T A)^I Y^T Z^T$$

using the identity

$$(Z^T A) (Z^T A)^{-1} = 1$$

yields

$$0 = A(Y^T Z^T A)^I (Y^T Z^T A) (Z^T A)^{-1} Z^T = AQ(Z^T A)^{-1} Z^T$$

where

$$Q = (Y^T Z^T A)^I (Y^T Z^T A)$$

is a projection operator.

This result is a special case of the following. A generalized inverse of a nonsingular matrix X is equal to the product of an idempotent operator and the true inverse

$$X^I = (X^I X)X^{-1} = X^{-1} X X^I$$

The left and right idempotent operators are the projectors $(X^I X)$ and (XX^I) respectively. If the Moore Penrose inverse is used the idempotent operators are the identity operators.

The projection operators developed for constrained least squares applications are also expressible in terms of 1, 2, 4 generalized inverses. The problem of minimizing $AX = b$ subject to $GX = 0$ had the solution $P_G b$ where

$$P_G = AQ(A^T A)^{-1} A^T$$

and

$$Q = 1 - (A^T A)^{-1} G^T [G(A^T A)^{-1} G^T]^{-1} G$$

Substituting $H = (A^T A^{-1}) G^T$ and $S^T = G$ yields $Q = 1 - H(S^T H)^{-1} S^T$ with the 1, 2, 4 generalized inverse $(S^T H^{-1}) S^T$.

4.8 Weighted Features and Subspace Methods

The incorporation of feature weighting into oblique projectors was considered. Justification for its use is provided below from a standpoint of least squares theory. The weighting of features is based on the statistical description of the measurement errors. The assumptions are that the average values of errors are zero and that the variances and covariances are known. If the model $R = b-Ax$ is adequate, the errors, R_i, associated with each row or measurement will be unbiased. By this is meant that with an ensemble of repeated measurement of b_i the set of R_i will have zero mean.[16] Using this ensemble the covariance matrix of the errors, Σ_R, can be formed. For obvious reasons the covariance matrix is often not known. In any event the estimate of the errors in the solution vector x is related to the covariance matrix, Σ_R. The ensemble of solution vectors $\{X_i\}$ has a covariance matrix

$$\Sigma_X = A^I \Sigma_R (A^I)^T \tag{1}$$

where A^I is the pseudoinverse of A. For unweighted least squares the inverse is the Moore Penrose inverse, $A^I = (A^TA)^{-1}A^T$. For weighted least squares the pseudoinverse can be written $A^I = (A^TWA)^{-1}A^TW$, where W is the weight matrix. The weight matrix must be positive definite. The covariance matrix for the solution vector, Σ_X is given by

$$\Sigma_X = (A^TWA)^{-1} \; (A^TW \, \Sigma_R \, WA) \; (A^TWA)^{-1} \tag{2}$$

It is the diagonal elements of Σ_X, $\sigma^2{}_{jj}$, that give the errors associated with X_j.

Gauss' Theorem states an important criteria for selection of weights. The theorem states that the weight which will give minimum variance is the inverse of the covariance of the measurement errors. That is if $W = \Sigma_R{}^{-1}$ then the $\sigma^2{}_{jj}$ will be a minimum. This intuitively makes sense as measurements with large variances will have small weights and measurements with small variances will be weighted heavily. The minimum variance covariance matrix for X is given by

$$\Sigma_X = (A^T\Sigma_R^{-1}A^T)^{-1} \tag{3}$$

If $\Sigma_R = \sigma^2 I$ then

$$\Sigma_X = \sigma^2(A^TA)^{-1} \tag{4}$$

Σ_R is usually not known and the assumption of it being a constant matrix $\sigma^2 I$ is ubiquitious since it shows no prejudice against any measurements. The unweighted least squares procedure will give an unbiased estimate of X and if the error covariance is a constant matrix, the unweighted least squares will give a minimum variance estimate of X. The penalty for using the wrong weight ($W \neq \Sigma_R^{-1}$) is the loss of minimum variance. A critical question concerning minimum variance is how sensitive is the variance to the wrong weight. (Note: no weight at all $W = 1$ is the wrong weight if Σ_R is not a constant matrix.) The selection of measurements or weighting of measurements is a standard approach in pattern recognition. The weighting or selection is based more on usefulness in distinguishing classes rather than on concerns about measurement error. The general rather than accidental success of such procedures would require that this type of measurement weighting does not have a large effect on the covariance. A simple example lends support to this idea. Consider a two class problem and a simple minimum distance classifier based on $b = Ax$ where A is m x 2 whose columns a_1 and a_2 are the average vectors for class 1 and class 2 respectively. The m features are measured with equal precision and the features are uncorrelated. The covariance matrix of the errors can be expressed as $\Sigma_R = \sigma^2 I$.

$$\text{Let } A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 1 \end{pmatrix} \qquad (5)$$

Even though the m measurements are of equal precision, only the first will be useful in distinguishing between the two classes. For the minimum distance classifier $X_1 > X_2$ if b belongs to class 1.

For this example

$$A^TA = \begin{pmatrix} m & m-2 \\ m-2 & m \end{pmatrix} \tag{6}$$

the condition number Cond(A^TA) = m-1 and the covariance can be calculated from (4) as

$$\Sigma_X = \sigma^2\left(A^TA\right)^{-1} = \frac{\sigma^2}{4(m-1)}\begin{pmatrix} m & 2-m \\ 2-m & m \end{pmatrix} \tag{7}$$

The uncertainties in X_1 and X_2 are obtained from the diagonal elements

$$\sigma_{ii}^2 = \frac{\sigma^2}{4}\left(\frac{m}{m-1}\right) \approx \frac{\sigma^2}{4}$$

and are independent of m for large m. The inclusion of a large number of precisely measured but unuseful measurements does not reduce σ_{ii} but does contribute to an increase in the condition number and to the correlation between class 1 and class 2.

A weighting scheme is used which attempts to minimize the apparent correlation between classes by minimizing the condition number. The following diagonal weight matrix will cause the apparent correlation to be zero.

$$w = \begin{pmatrix} 1 & & & & 0 \\ & \frac{1}{m-1} & & & \\ & & \frac{1}{m-1} & & \\ & & & \ddots & \\ 0 & & & & \frac{1}{m-1} \end{pmatrix}$$

w is a mxm diagonal matrix. The model now uses a weighted least squares calculation. The matrix (A^TWA) is diagonal and class 1 appears uncorrelated with class 2. The condition number $\text{Cond}(A^TWA) = 1$, an absolute minimum. We have

$$(A^TWA) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad (A^TWA)^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

and

$$\begin{matrix} X_1 \\ X_2 \end{matrix} = (A^TWA)^{-1} A^TWb$$

Since these weights are not the inverse of the covariance of the measurement errors, the covariance of X must be calculated from Eq. (2) rather than Eq. (3) and in principle will not yield the minimum variances. However substitution of the required quantities in Eq. (2) yields the same estimate for Σ_X as was obtained in Eq. (7). The use of the incorrect weights in this model has not had an adverse effect on the solution variances. The weighting generates a numerically stable model with a well conditioned normal matrix which is easily inverted. This example offers some hope for the use of weighting schemes which minimize apparent correlation or minimize the importance of measurements which are not useful for classification.

The use of weights in oblique projection techniques is attractive. The weighted projection method can be formulated so that once the projector is generated during a learning step, no additional burden will be incurred. The classification step is as simple as in the unweighted case. The form of the oblique projection operator used here is $O_A = A(Z^TA)^IZ^T$ where Z is a basis for the projection onto A which is orthogonal to ϕ_B. That is $Z^TB = 0$ for any vector B that belongs to ϕ_B. When using weighted measurements a $\hat{Z}$

is sought which is W orthogonal to ϕ_B. In the unweighted case Z is constructed from

$$Z = A - B(B^T B)^{-1} B^T A$$

For the weighted case the desired $\hat{Z}$ is

$$\hat{Z} = WA - WB(B^T WB)^{-1} (A^T WB)$$

Then

$$\hat{Z}^T B = A^T WB - (B^T WB) (B^T WB)^{-1} (A^T WB) = 0$$

and the weighted projection is given by

$$\hat{O}_A = A(\hat{Z}^T A)^I \hat{Z}^T$$

$\hat{O}_A$ works directly on unweighted vectors. Thus no additional burden over unweighted projections occurs once $\hat{Z}$ and $\hat{O}$ are formed.

References for Section 4

1. S. Watanake, P.F. Lambert, C.A. Kulikowski, J.L. Buxton, and R. Walker, "Evaluation and Selection of Variables In Pattern Recognition," Computer and Inform. Sciences, Vol. 2, (J. Tou ed.) p. 91, Academic Press, NY (1967).

2. C.A. Kulikowski and S. Watanabe, "Multiclass Subspace Methods in Pattern Recognition," in Proc. Nat. Electron Conf., Chicago, IL (1970).

3. S. Watanake and N. Pakvasa, "Subspace Method in Pattern Recognition," Proc. 1st Int. Joint Conf. Pattern Recognition, Washington, DC (1973).

4. C.W. Therrien, "Eigenvalue Properties of Projection Operators and Their Application To the Subspace Method of Feature Extraction," IEEE Trans. on Comput. C-24, 944 (1975).

5. Svante Wold, "Pattern Recognition By Means of Disjoint Principal Components Models," Pattern Recognition 8, 127 (1976).

6. Michael Sjöström and Svante Wold, "SIMCA: A Pattern Recognition Method Based on Principal Component Models," Pattern Recognition in Practice, E.S. Gelsema and L.N. Kanal (eds.) North Holland Publishing Company (1980).

7. T. Kohonen, "Associative Memory - A System Theoretical Approach," Springer-Verlag Berlin (1979).

8. E. Oja and J. Karhunen, "An Analysis of Convergence for a Learning Version of the Subspace Method,: " J. Math. Anal. Appl. 91, 102 (1983).

9. E. Oja, "Subspace Methods of Pattern Recognition," Research Studies Press Ltd, John Wiley (1983).

10. Y. Noguchi, "Subspace Method and Projection Operators," Proc. 4th Int. Conf. on Pattern Recognition, p. 449 Kyoto (1978).

11. S.N. Afriat, "Orthogonal and Oblique Projectors and the Characteristics of Pairs of Vector Spaces," Proc. Cambridge Philos. Soc. 53, 800 (1957).

12. R.D. Milne, "An Oblique Matrix Pseudoinverse," SIAM J. Appl. Math 16, 931 (1968).

13. T.N.E. Greville, "Solutions of the Matrix Equation XAX = X and Relations Between Oblique and Orthogonal Projectors," SIAM J. Appl. Math 26, 828 (1974).

14. Adi Ben-Israel and T.N.E. Greville, "Generalized Inverses: Theory and Application," John Wiley (1974).

15. David G. Luenberger, "Introduction to Linear and Nonlinear Programming," Addison-Weseley (1973).

16. W.R. Draper and H. Smith, "Applied Regression Analysis," Second Ed., John Wiley (1981).

END

DTIC

6-86